# АВТОМАТИЗИРАНО ГЕНЕРИРАНЕ НА МЕТАДАННИ ЗА УЧЕБНИ ОБЕКТИ

***Христина Костадинова, Георги Тотков, Димитър Благоев***

*ЮЗУ „Неофит Рилски”, ПУ „Паисий Хилендарски”*
*kostadinova@swu.bg, {totkov, gefix}@uni-plovdiv.bg*

***Резюме:*** *В работата са разгледани подходи за автоматизирано генериране на метаданни за учебни обекти. За основа е приета схемата на стандарта LOM, в която за всяко поле с метаданни са предложени подходящи методи за извличане. Всички методи са групирани в пет основни категории, като особено внимание се обръща на първата от тях, която използва отговори на т. нар. ‘акумулативни въпроси’ при е-обучение. Експериментирането на различни методи за извличане на метаданни (напр. в самия процес на е-обучение) е първата стъпка към създаване на автоматизирана система за генериране на метаданни в процеса на е-обучение.*

***Ключови думи:*** *извличане на метаданни, LOM, акумулативни въпроси*

## 1. Въведение

С навлизането на е-обучението се появява необходимостта от създаване на софтуерни системи за автоматизирано генериране на електронни курсове. Изключително важно за всеки такъв курс е описанието на характеристиките на включените в него учебни обекти (LO - Learning objects), т.е. създаването и/или генерирането на метаданните.

Основен стандарт в областта на метаданните за обекти, използвани в обучението е LOM (Learning Object Metadata) [7], според който метаданните – на брой 58, са разпределени в девет групи: 1. Основни (General), 2. Продължителност на съществуване (Life Cycle), 3. Данни за метаданните (Meta-Metadata), 4. Технически характеристики (Technical), 5. Образователни (Educational), 6. Права (Rights), 7. Връзки (Relation), 8. Анотация (Annotation) и 9. Класификация (Classification). Всяка група съдържа определен брой полета, които я характеризират в детайли. Освен че описват съдържанието на учебния обект, метаданните се отнасят и до други елементи на процеса на обучение. С използване на метаданни, аташирани към съответните учебни обекти, се улеснява тяхното търсене и обмен, като значително се повишава качеството на администриране на системите за е-обучение. От друга страна, създаването и въвеждането на метаданни е трудоемка и високо квалифицирана експертна дейност, което налага търсене на различни начини за тяхното автоматизирано генериране.

Според [8], автоматизираното генериране на метаданни за даден учебен обект е резултат от машинна обработка, в която участват единствено

разработчик, разпространител на софтуер и инициатор на процеса. Последното е валидно само при условие, че метаданните за даден LO са статични и неизменяеми при участие на дадения обект в различни процеси на виртуално обучение. Подобно, между другото, е и разбирането на повечето автори, които предполагат (явно или не), че метаданните, съпровождащи учебните обекти (вкл. тестовите единици), са относително неизменни – независимо от процесите, в които участват. Подобна предпоставка силно стеснява възможностите за удовлетворително решаване на проблема за автоматизирано генериране на метаданни за LO. Например, генерирането на метаданни, свързани с трудността, областта, степента на гранулираност, типична възраст на обучаемите, средното време, необходимо за усвояване на материала, и др. под. на даден LO, няма как да не зависи силно от конкретния контекст на провежданото обучение. Нещо повече, динамичният характер на редица метаданни за LO (напр. трудност), едва ли може да бъде поставен под съмнение (напр. извън зависимост от постиженията на конкретна група обучавани). Очевидно, че за решаване на посочения проблем, трябва да бъдат изследвани и други подходи.

## 2. Подходи за автоматично и автоматизирано генериране на метаданни

Съществуват различни методи за автоматично и автоматизирано извличане на метаданни за даден LO в обучението. По-голямата част от тях се основават на стандарта LOM − Learning Objects Metadata (вж. например [3, 5, 13]). Разработени са и гъвкави инструменти, които не следват известни стандарти, а разглеждат връзката на потребителите с учебните обекти при поставена конкретна цел [10].

В системите за автоматизирано генериране на метаданни се използват методи за извличане, които зависят от типа на съответната метаданна (низ, число, елемент от списък, и т.н.) и от вида на съответния LO.

Обикновено се използват ***два основни подхода*** за генериране на метаданни − съответно базирани на ресурси или на контекста на материала [3, 12].

Първият подход е комбинация от следните методи:

- *добиване* (harvest) на метаданни чрез събиране на вече създадени метаданни от съществуващи хранилища;
- *извличане* (extraction) – необходимата информация се открива в съдържанието на материала (напр. търсене на ключови думи);
- *класификация* –определяне на стойностите на метаданните от специално съставени речници, например идентификатори на езици;
- *разпространяване (*или наследяване) с използване на връзка на конкретния материал с останалите, и неговата позиция в структурата на курса.

В зависимост от конкретния материал и хранилището, в което се извършва търсенето, дадена метаданна може да бъде генерирана чрез повече от един метод. Съществуват и метаданни, които не могат да бъдат определени чрез нито един от тези подходи. За тях е възможно да се приложи вторият (контекстно базиран) подход.

Метаданните могат да бъдат извличани от различни източници. Например в [3] са разграничени пет подобни източника: системата за управление на обучението (LMS − Learning Management System) и използваната операционна система; log файловете на сесиите; профилите на потребителите; обратната връзка - информацията, която се получава в следствие на работата на потребителите с материалите по време на обучение (напр. време, необходимо за усвояване на материала, резултати от изпитвания и др).; други LO и техните метаданни.

В [5] се изследва подход, който анализира контекста на документа и включва подробно изследване на профила на неговия създател и на информацията за курса, в който е включен материала. Системата за управление на изучаваното съдържание дава информация относно връзките между обектите, начинът за използване на метаданните за материали, които съдържат даден LO. В системата за управление на е-обучението се съдържа богата контекстуална информация (напр. в кои курсове се използва даден обект, колко пъти е осъществяван достъп до него, колко пъти е ‘свалян’ и др.), която също може да бъде използвана за определяне на метаданни. Едновременно с концептуалния анализ се използват и други методи: анализ на съдържанието на документа (включващо езикови идентификатори, методи за извличане на ключови думи, разпознаване на образи за изображенията); оценка на данни, свързани с използването на документа в системата (напр. време, проведено в изучаване на даден материал; време за решаване на определени задачи и т. н.); анализ на информацията, свързана със сложността на структурата на документа (напр. някои материали − част от учебния обект, понякога се съхраняват отделно и са източник на данни за останалите), и др.

Подобни подходи за генериране на метаданни е възможно да се осъществят в рамките на конкретни среди за обучение (напр. в среда за обучение *AdaptWeb* [14]). В основата на реализация [13] е стандартът LOM, като в зависимост от вида на съответните метаданни, процесът на генериране се извършва автоматично или автоматизирано. За автоматично генериране на метаданни се използват 4 (четири) метода:

- анализиране на физическия файл, който съдържа учебния материал и извличане на данни за попълване на метаданни от раздел 4. Технически характеристики на LOM;
- търсене и извличане на метаданни от базата данни с потребители, курсове и материали на системата за е-обучение (в сл. AdaptWeb);

- търсене и извличане на метаданни от съответната БД за използваните LO;
- анализиране на XML-документи от БД, извличане и складиране на данни с използване на съответните тагове.

При втория (автоматизиран) подход се предполага диалог с потребителя за потвърждаване и уточняване на данните, получени при автоматично извличане. Използват се два метода:

- определяне на стойности по подразбиране (предположение за възможните метаданни);
- търсене по шаблон на обект, подобен на изследвания (за целта е необходимо шаблонът да е предварително създаден от потребители в XML базата данни).

Общото и в трите подхода е, че са базирани на стандарта LOM (главно − поради неговата популярност в света на метаданните за учебни обекти, използвани в е-обучението). Извличането на метаданни от физическия файл, който съдържа материала, търсенето в хранилища с метаданни, разглеждането на връзките между обектите и тяхното взаимодействие с потребителите, са основните подходи, които се срещат в представените системи. В изследване [3], освен по-горе посочените методи е експериментиран и подход, който използва отговори на потребителите на софтуер за електронни курсове. Опитът е направен за метаданни ‘ключови думи’ (поле 1.5 на LOM) и постига добри резултати - около 90% от получените отговори съвпадат с думите и фразите, посочени от експерта, който въвежда материала в системата.

В настоящата работа се изследва подход, свързан с хипотезата, че определени метаданни за LO − в по-малка или в по-голяма степен, зависят и се определят от съдържанието, историята, субектите и събитията на процеси на обучение, в които участва дадения LO. В случая на автоматизирано генериране на тестови единици и съответни метаданни, първите изследвания в указаната посока, са проведени в [11] с въвеждане на т. нар. ‘акумулативни тестови единици’.

Целта тук е - да се изследва приложимостта на подобни ‘акумулативни’ подходи за автоматично и автоматизирано генериране на метаданни, свързани с LOM.

### 3. Към автоматизирана система за генериране на метаданни

В Табл. 1. е представена класификация на методите за автоматично (или автоматизирано) извличане на метаданни от LO. Включването на даден метод към някоя от 6-те групи се определя от неговата същност – основният принцип, на който се основава неговото прилагане.

| Група | Същност | Знак |
|---|---|---|
| 1. | Акумулативни тестови въпроси (test items), генерирани при виртуално обучение | ● |
| 2. | Автоматизирано извличане от текстове | ♦ |
| 3. | Анализ на връзки и взаимодействия | ۞ |
| 4. | Търсене в хранилище с метаданни (при въвеждане) | □ |
| 5. | Информация за самата система | ▲ |
| 6. | Все още неопределена | ? |

**Таблица 1. Групи от методи за извличане на метаданни по стандарт LOM**

В Табл. 2. се представя схемата от метаданни на стандарта LOM, като за всяка метаданна (поле в съответната таблица) се посочва групата на реализиран и експериментиран метод за нейното генериране.

| № | Име | Група | № | Име | Група |
|---|---|---|---|---|---|
| **1** | **General** | | 4.6 | Other Platform Requirements | ? |
| **1.1** | Identifier | | 4.7 | Duration | ♦ |
| **1.1.1** | Catalog | □ | **5** | **Educational** | |
| **1.1.2** | Entry | □ | 5.1 | Interactivity Type | ▲ |
| **1.2** | Title | ●□ | 5.2 | Learning Resource Type | ▲ |
| **1.3** | Language | ♦ | 5.3 | Interactivity Level | ♦ |
| **1.4** | Description | ●□ | 5.4 | Semantic Density | ♦□ |
| **1.5** | Keywords | ●♦□ | 5.5 | Intended End User Role | □۞ |
| **1.6** | Coverage | ●□ | 5.6 | Context | □۞ |
| **1.7** | Structure | ●□ | 5.7 | Typical Age Range | □۞ |
| **1.8** | Aggregation Level | ●□ | 5.8 | Difficulty | □۞ |
| **2** | **Life Cycle** | | 5.9 | Typical Learning Time | ♦ |
| **2.1** | Version | ۞ | 5.10 | Description | □ |
| **2.2** | Status | ♦ | 5.11 | Language | ♦ |
| **2.3** | Contribute | | **6** | **Rights** | |
| **2.3.1** | Role | □۞ | 6.1 | Cost | □ |
| **2.3.2** | Entity | □۞ | 6.2 | Copyright and Other Restrictions | ♦□ |
| **2.3.3** | Date | ♦ | 6.3 | Description | □ |

| 3 | Meta-Metadata | | 7 | Relation | |
|---|---|---|---|---|---|
| **3.1** | Identifier | | 7.1 | Kind | □⚙ |
| **3.1.1** | Catalog | ▲ | 7.2 | Resource | |
| **3.1.2** | Entry | ▲ | 7.2.1 | Identifier | |
| **3.2** | Contribute | | 7.2.1.1 | Catalog | □⚙ |
| **3.2.1** | Role | □⚙ | 7.2.1.2 | Entry | □⚙ |
| **3.2.2** | Entity | □⚙ | 7.2.2 | Description | ● |
| **3.2.3** | Date | ♦ | **8** | **Annotation** | |
| 3.3 | Metadata Scheme | ▲ | 8.1 | Entity | □ |
| 3.4 | Language | ♦ | 8.2 | Date | ♦ |
| **4** | **Technical** | | 8.3 | Description | □ |
| 4.1 | Format | ♦ | **9** | **Classification** | |
| 4.2 | Size | ♦ | 9.1 | Purpose | □ |
| 4.3 | Location | ♦ | 9.2 | Taxon Path | □⚙ |
| 4.4 | Requirements | | 9.2.1 | Source | □ |
| 4.4.1 | OrComposite | | 9.2.2 | Taxon | □⚙ |
| 4.4.1.1 | Type | ♦ | 9.2.2.1 | Id | □ |
| 4.4.1.2 | Name | ♦ | 9.2.2.2 | Entry | □ |
| 4.4.1.3 | Minimum Version | ? | 9.3 | Description | ● |
| 4.4.1.4 | Maximum Version | ? | 9.4 | Keywords | ● |
| 4.5 | Installation Remarks | ? | | | |

**Таблица 2. Метаданни по LOM и подходящи (за извличане) групи методи**

***Първата група методи*** е свързана с използване на системи от акумулативни въпроси, подходящи за определен вид метаданни, свързани със съответния LO (текст, тестова единица, събитие и т.н.), а така също и с други LO, използвани многократно в контекста на различни процеси на виртуално обучение. Примери на акумулативни въпроси, отнасящи се до метаданни от схемата на LOM, са представени в Табл. 3. Елемент на подобно обучение могат да бъдат не само обучавани, но и преподаватели – експерти в изучаваната предметна област (ПО), като отговорите на последните могат да имат по-съществен принос (напр. по-голяма тежест) в процеса на генериране. Следвайки принципите [11, 12] на акумулиране на отговори и генериране на нови типове тестови единици (вкл. с използване на схеми за оценяване на различни когнитивни равнища [2]), в процеса на виртуално обучение могат да

бъдат постоянно натрупвани и оценявани кандидат-данни за този вид метаданни.

| Метаданна | Примерни акумулативни въпроси |
|---|---|
| 1.2 Title | Посочете най-подходящото (според Вас) заглавие на учебния обект |
| 1.4 Description | Резюмирайте (до 200 знака) материала |
| 1.5 Keywords | Избройте (поне 5) ключови понятия, свързани с учебния обект. |
| 1.6 Coverage | Посочете периода, за който се отнася материала. Посочете (поне 3) основни тематични области, свързани с материала (желателно в йерархичен ред) |

**Таблица 3. Примерни акумулативни въпроси по стандарт LOM**

***Втората група методи*** (за автоматизирано извличане) включва − от една страна, подходи за търсене на информация в текстовото съдържание на учебния материал, и извличане на данни от физическия файл, в който е съхранен (напр. попълване на технически характеристики според данните във визитката на файла); от друга − методи за класификация спрямо предварително създадени речници, например идентификация на езика.

В [1] е разгледан метод за извличане на данни посредством специализирани граматики и регулярни изрази. Използваният подход е подходящ за извличането на метаданни, които да с достатъчно ясна и отличаваща ги структура и не изискват сложен контекст за тяхната коректна идентификация. Този подход често дава най-добри резултати за данни, които лесно могат да бъдат отделени и извлечени от текста, но изисква значителни умения за реализирането на необходимите граматики за тяхното извличане. От друга страна методите на принципа на обучението върху входни данни [4] са по-лесни за използване за сметка на допълнителния процес на натрупване и анализиране на обучаващи примерни метаданни.

Методите от ***третата група*** провеждат анализ на връзки, отношения, взаимодействия и данни за потребители, структура на обекти и др. под. в автоматизираната система.

Методите от ***четвъртата група*** са предназначени за търсене и откриване на съответни метаданни във вече създадени хранилища. В [6] е направено проучване на такива хранилища, базирани на LOM. Изследвани са приликите и разликите в подходите за изграждане на такива структури. На първо място е необходимо да бъде определен стандартът на метаданните, използвани в съответната автоматизирана система. Решението на този въпрос зависи най-вече от съдържанието и предназначението на ползваните

електронни документи. По принцип е възможно и ‘смесване’ на елементи от различни стандарти.

***Петата група*** от методи определят метаданни, които се формират още при създаването на системата за обучение и схемата за мета-метаданни.

**4. Заключение**

Целта на разработката е да систематизират и изследват подходящи подходи за извличане на метаданни за LO. За провеждане на експерименти е избран стандарта LOM, поради неговото широко използване и факта, че ще се изследват материали от различни области. Проучването на различни методи за автоматизирано извличане на метаданни показва, че съществуват добри възможности за минимално участие на автора на LOs в съставянето на визитка на учебния обект. Както беше посочено, може да се разграничат 5 (пет) основни групи подходи. Изследването на методи от първата група (акумулативни въпроси, генерирани в резултат на виртуално обучение с различни групи обучавани) се оказва особено продуктивно, и ще бъде предмет на следващи публикации. Особено предизвикателна е задачата по изграждане на курсове за е-обучение ‘от нулата’. Последното означава – стартиране на виртуално обучение от набор LO, свързани със съответната предметна област (дори при отсъствие на съпровождащи метаданни и банка от тестови въпроси), и постепенното генериране (на базата на подходящи акумулативни тестови единици и на метаданни за всеки използван в обучението LO, и на тестови въпроси от различен тип (на базата на естествени дистрактори – резултат от ‘колекционирани’ и оценени отговори на обучаваните в резултат на зададените акумулативни въпроси). Перспективно направление на изследванията в областта на интелигентните системи за е-обучение е свързано с развитие на една друга идея − включване в процеса на виртуално обучение и на акумулативни въпроси, свързани с когнитивни равнища на знанието (напр. по таксономията на Блум).

Разработката е частично финансирана по проект ДО 02-308 към Националния научен фонд и проект BG 051PO001-3.3.04/13 на ОП „Развитие на човешките ресурси“ на Европейския Социален Фонд (2007−2013).

**ЛИТЕРАТУРА**

[1] Благоев Д., *Извличане на контактна информация от документи на български език*, в Сборник доклади на 3-та научна конференция за студенти, докторанти и млади научни работници, 25.4.2009, Пловдив, 339-343.

[2] Костадинова Хр., Г. Тотков, М. Райкова, *Към автоматизирано генериране на тестове по Блум*, 40-та Юбилейна конференция на СМБ, Боровец, 5 – 9 април 2011 г., 413-422.

[3] Bauer M., R. Maier, *Metadata Generation for Learning Objects: An Experimental Comparison of Automatic and Collaborative Solutions*. E-Learning, 2010, pages181-195.

[4] Blagoev D., G. Totkov, M. Staneva, Kr. Ivanova, Kr. Markov, *Indirect Spatial Data Extraction from Web Documents*, International Book Series Information Science & Computing, New Trends in Intelligent Technology, Number 14, ITHEA, 2009, 89-100.

[5] Cardinaels K., M. Meire, E. Duval, *Automating Metadata Generation: the Simple Indexing Interface*. In Proceedings of the 14th international conference on World Wide Web, ACM Press, 2005. 548-556.

[6] Duval, E., F. Neven, *Reusable Learning Objects: a Survey of LOM-based Repositories*. MULTIMEDIA ’02 Proceedings of the tenth ACM international conference on Multimedia, 2002.

[7] *Final LOM Draft Standard*, http://ltsc.ieee.org/wg12/20020612-Final-LOM-Draft.html.

[8] Greenberg, J., K. Spurgin, A. Crystal, *Functionalities for Automatic-Metadata Generation Applications: A Survey of Metadata Experts’ Opinions*. International Journal of Metadata, Semantics, and Ontologies. 2006, *Vol. 1, No. 1, 2006* 3 http://www.inderscience.com/storage/f121932106117458.pdf.

[9] Handschuh S., S. Staab, F. Ciravegna, *S-CREAM - Semi-automatic CREAtion of Metadata*. EKAW '02 Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web 2002.

[10] McCalla, G., C. Brooks, *Towards Flexible Learning Object Metadata*. Engineering Education and Lifelong Learning, (16):50–63, 2006.

[11] Sokolova M., G. Totkov, Accumulative Question Types in e-Learning Environment, International Conference on Computer Systems and Technologies - CompSysTech’2007, IV.21-1 - IV.21-6.

[12] Sokolova, M., G. Totkov, Extended IMS Specification for Accumulative Test System, ACM International Conference Proceeding Series; Vol. 374, Proc. of the 9th Int. Conf. on CompSysTech, Gabrovo, Bulgaria, June 12-13, 2008, V.14-1–V.14-6.

[13] Warpechowski M., M. A. M. Souto, J. P. M. de Oliveira, Techniques for Metadata Retrieval of Learning Objects. Workshop on Applications of Semantic Web Technologies for e-Learning (SW-EL@AH'06), 2006.

[14] Warpechowski M., P. de Oliveira et al., Adaptive Hypermedia in the AdaptWeb Environment In: First EAW. Eindhoven. In: AH 2004 Workshop Proceedings, 68 – 73.